Globs in the Primordial Soup

The Emergence of Connected Crowds in Mobile Wireless Networks

Simon Heimlicher
Computer Engineering and Networks Laboratory
ETH Zurich, Switzerland
heimlicher@tik.ee.ethz.ch

Kavé Salamatian
LISTIC PolyTech
Université de Savoie Chambery Annecy, France
kave.salamatian@univ-savoie.fr

ABSTRACT

The analysis of mobile networks spurred a lot of research and recently focused on node encounters, or "contacts", as the contact process determines whether space-time path connecting far-flung nodes exist. But the contact process alone yields an incomplete picture. People tend to gather at points of interest and as a consequence, connected components (clusters) arise. Nodes may be scattered across a large area, but many practical scenarios feature sizable clusters, sometimes comprising the majority of nodes. In this paper, we analyze the distribution of cluster size in mobile networks. Specifically, we model stochastic coalescence (merging) and fragmentation (splitting) of clusters as a Markov chain and derive analytically the exact stationary distribution of cluster size. Moreover, we prove that as the number of nodes grows, the clustering behavior converges to a mean field. The mean field is obtained as a closed-form expression and is a surprisingly good approximation for practical scenarios. What is remarkable about this result is that the mean field translates a microscopic property — the parameters of the merge–split process — to the cluster size distribution, an important macroscopic property of the system. Thus, our model allows characterizing systems of mobile nodes through the parameters of the merge–split process that yield the empirically observed cluster size distribution. We validate the mean field approximation and the exact distribution against random walk simulation as well as real mobility traces with sizes spanning three orders of magnitude and ranging from conference visitors to taxicabs.

1. INTRODUCTION

Up to now, the analysis of mobile networks predominantly modeled individual nodes; in particular research on delay-tolerant networking (DTN) focused on characterizing the process governing single-hop paths ("contacts") and leveraging the resulting "space-time paths" required for communication in a disconnected network. Those studies laid an important foundation toward understanding the contact process. Related work evaluating contact-based routing and forwarding schemes reports promising results for novel applications specifically designed for those networks. In addition to space-time paths, we argue that multi-hops paths may exist, especially if future mobile network scenarios become larger and denser. Indeed, multi-hop paths might allow running some of the applications we use every day in a mobile wireless network even though those applications are not designed to be as delay-tolerant as those that have traditionally been studied in the DTN community. We use the term *partial path* to refer to a chain of connected nodes bridging part of the gap between source and destination, as opposed to the term *full path*, which is a path from source to destination.

Following those definitions, several questions arise naturally: do partial paths exist? Continuum percolation theory (see [8] and references therein) would have you believe otherwise; its main result is that a phase transition occurs between the disconnected and the connected regime; *i.e.*, if the node density is below the percolation threshold, almost all nodes are isolated; above the threshold, the network "percolates" and forms one giant connected component. But the caveat is that this only applies if node distribution is uniform, thus excluding the vast majority of real-world scenarios. In delay-tolerant networking, studies of space-time paths [6, 10] as well as mobility models featuring clustering behavior abound [22, 18, 20]; yet those works are concerned with metrics of relevance to contact-based forwarding schemes and largely skip over the existence of multi-hop paths.

In light of this situation, we argue that a methodology specifically for modeling partially-connected scenarios is in order. A major roadblock on the way toward modeling partial paths is that by definition a path involves multiple nodes and implies dependency between those nodes. Observing that individual node positions are irrelevant for analyzing existence and properties of partial paths, we could model partial paths instead of nodes. Indeed, we can go even further and model all partial paths that are lumped together in a connected component (cluster) as a whole. According to this approach, a network is described by the dynamics of cluster creation and disappearance and the transient and stationary distribution of the cardinality of those clusters.

We propose a model based on the concept of stochastic coagulation and fragmentation in particle systems. More specifically, we model a system of mobile nodes in analogy to a system of particles in a solvent. In analogy to how those globs of particles coalesce and fragment, we describe a sys-

tem of N mobile nodes as a set of clusters; the state of the system is represented by a vector in which every element $i = 1, 2, \ldots$ represents the number of clusters of size i. Assuming that N is constant, there are two primitive events that can happen in this system. First, two clusters of sizes k and l can merge into a new cluster of size $k + l$ as described by the merge process. Second, the inverse can happen as well, *i.e.*, a cluster of size $k+l$ splitting into two clusters according to the split process. The merge and the split process determine the stationary distribution of cluster size and thus, whether the network is connected, disconnected, or partially connected, the latter implying existence of partial paths.

We implement this model as a Markov process over the finite state space of all partitions of N. Under certain conditions, this process is reversible and thus its stationary distribution (corresponding to the distribution of cluster size) is obtained in closed form. Furthermore, we prove that the behavior of the merge–split process converges to a mean field for large numbers of nodes, N. The mean field is obtained in closed form and even for realistic networks with finite numbers of nodes provides a very useful approximation.

While the focus of this paper is clearly on studying in depth the merge–split model, we also aim to illustrate its usefulness to the extent space permits. The mean field approximation intuitively translates through a simple expression a microscopic property——the parameters of the merge-split process——to the cluster size distribution. Thus, our analysis allows characterizing systems of mobile nodes solely by the merge/split parameters that yield the observed cluster size distribution. The number of non-singleton clusters indicates, how well contact-based or partial-paths-based algorithms fit a given scenario.

The mean field approximation as well as the exact distribution are validated against random walk simulation as well as mobility traces. We use traces from conference visitors and taxicabs in San Francisco and Shanghai. More specifically, we extract the merge and the split rate from these scenarios and then derive the cluster size distribution using both our exact analytical solution and the mean field approximation. The exact result yields a remarkably precise prediction; the mean field approximation by its nature mainly gives a reliable prediction of the shape of the distribution; in particular it predicts whether giant components emerge.

There are several other applications of our model, such as predicting the distribution of the time to forward messages between nodes; we are currently working on those results and they are to be published in a forthcoming paper. Below, we summarize the contributions we present in this paper.

- We introduce a a merge–split process modeling the stationary distribution of cluster size in mobile networks;
- we prove convergence to a mean field behavior with increasing number of nodes, thus providing a closed-form expression to translate the microscopic behavior (merge–split process with three parameters) to the macroscopic behavior (cluster size distribution);
- we validate the predicted cluster size distribution (exact derivation and mean field approximation) against random walk mobility and three real-world traces.

In the next section, we introduce the merge–split model, we prove its convergence to a mean field, and we outline calibration of the model with empirical data. We validate this model and the calibration in Sec. 3, where we calibrate the model with synthetic as well as three real-world mobility traces and compare the predicted cluster size distribution with the empirical one. Section 4 discusses our contributions vis-a-vis related work and Sec. 5 concludes and outlines future work.

2. ANALYTIC FORMULATION

2.1 Finite size system formulation

We describe an arbitrary mobile network as a system of N interacting nodes. At every time t, a node is in exactly one cluster, *i.e.*, it is member of a set of nodes connected by a full path at time t. The state of the system is described by the cluster size vector $(\nu_N(1,t), \nu_N(2,t), \ldots, \nu_N(N,t))$ with elements $\nu_N(i,t)$ representing the number of clusters of size i at time t. We consider two primitive interactions between these nodes.

1. **Merge reaction:** A cluster of k nodes merges with a cluster of l nodes, yielding a cluster of $k + l$ nodes:

   $$C_k + C_l \to C_{k+l}.$$

   This reaction is also called *coalescence* and happens at a rate $\psi_N(k,l)$, which is assumed to be symmetric, *i.e.*, $\psi_N(k,l) = \psi_N(l,k)$. A merge reaction of clusters of sizes k and l has the following drift effect on the cluster size vector: $(\ldots, \nu_N(k,t) - 1, \ldots, \nu_N(l,t) - 1, \ldots, \nu_N(k+l,t) + 1, \ldots)$.

2. **Split reaction:** A cluster of size l splits into two clusters of sizes k, $(k < l)$ and $l - k$:

   $$C_l \to C_k + C_{l-k}.$$

   This reaction is also called *fragmentation* and happens at a rate $\phi_N(l|k)$ and we assume $\phi_N(l|k) = \phi_N(l|l-k)$. A split reaction has the following drift effect on the cluster size vector: $(\ldots, \nu_N(l-k,t)+1, \ldots, \nu_N(k,t)+1, \ldots, \nu_N(l,t) - 1, \ldots)$.

We call such a process a merge–split process. These reactions happens subject to the node conservation condition:

$$\sum_{k=1}^{N} k\nu_N(k,t) = N, \forall t \leq 0. \tag{1}$$

This defines a Markov process over the finite state space $\Omega = \Omega_N = \{\tau\}$ of all partitions of N. A special case of this process with only the merge reaction is called Marcus-Lushnikov process [17, 16] and has gained attention from the

mathematical community. The analogous process involving only the split reaction is called fragmentation process and has been studied extensively in the context of branching processes. The problem we analyze here is a mix of these two problems.

For ease of notation, we will drop index N referring to the total number of nodes in the forthcoming unless needed. As we will see it is useful to define the following intensity ratio function $q(k,l)$, based on the ratio between merge and split intensity, as:

$$q(k,l) = \begin{cases} \dfrac{\psi(k,l)}{\phi(k+l|l)}, & \text{if } \psi(k,l)\phi(k+l|l) \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

The existence of a stationary equilibrium state is conditioned on the reversibility of the Markov chain; a Markov process $\mathcal{M}_t$ is said to be reversible with respect to a probability measure μ if for all $t \geq 0$, the process $\mathcal{M}_t^{\mu}$, $0 \leq s \leq t$ and $\mathcal{M}_{t-s}^{\mu}$, $0 \leq s \leq t$, starting from the same initial distribution μ, have the same finite dimensional distribution [13]. Reversibility is an important property of a Markov process; if a reversible process is ergodic, its unique stationary distribution is the reversible measure. The reversible measure can be derived in general using the flow equilibrium equation of the Markov chain, *i.e.*, $V(\tau,\xi)\mu(\tau) = V(\xi,\tau)\mu(\xi)$, where $V(\tau,\xi)$ is the total intensity of transitions from state τ to ξ.

The following theorem gives necessary and sufficient conditions under which the process is reversible and will therefore converge to a stationary equilibrium state. Note that we only include the intuition of the proofs of the theorems given in the forthcoming due to lack of space; the complete proofs are provided in [2].

THEOREM 1. *[13] Suppose that $q(k,l) > 0$, for $2 < k+l < N$, then the merge–split Markov process is reversible if and only if for some function $a(k) > 0,\ k = 1,\ldots,N$, we can rewrite $q(k,l)$ as*

$$q(k,l) = \frac{a(k+l)}{a(k)a(l)}.$$

Moreover a merge–split process will have a reversible distribution μ following the closed form formula derived as:

THEOREM 2. *Suppose that $q(k,l)$ follows the condition given in Thm. 1, then the merge–split process defined above is reversible with respect to the invariant measure $\mu = \mu_N \in \Omega_N$, given by*

$$\mu_N(\tau) = \mathcal{C}_N \frac{a(1)^{n_1} a(2)^{n_2} \ldots a(N)^{n_N}}{n_1! n_2! \ldots n_N!}. \tag{3}$$

where $\tau(t) = (n_1,\ldots,n_N) \in \Omega_N$ is an acceptable configuration with n_k clusters of size k. $\mathcal{C}_N$ is a scaling coefficient defined such that $\sum_{\tau\in\Omega_N} \mu(\tau) = 1$.

The proof of the theorem proceeds by validating that this distribution satisfies the flow equilibrium condition, $V(\tau,\xi)\mu(\tau) = V(\xi,\tau)\mu(\xi)$, $\tau,\xi \in \Omega_N$. In the forthcoming we will denote

$$c_N = \frac{1}{\mathcal{C}_N} = \sum_{\tau\in\Omega_N} \frac{a(1)^{n_1} a(2)^{n_2} \ldots a(N)^{n_N}}{n_1! n_2! \ldots n_N!}. \tag{4}$$

The invariant measure gives the stationary state occupation measure, *i.e.*, the probability that the Markov chain is in state τ in equilibrium. However, for our purpose, we are interested in knowing the statistics of $n_k(\tau)$, *i.e.*, the number of clusters of size k in the configuration τ. The below statistics are of interest:

$$\begin{aligned} \nu_N(k) &= \mathbb{E}\left\{n_k(\tau)\right\}, && k = 1,\ldots N \\ \varsigma_N(k,l) &= \mathbb{C}\text{ov}\left\{n_k(\tau), n_l(\tau)\right\}, && k \neq l = 1,\ldots,N \\ \sigma_N^2(k) &= \mathbb{V}\text{ar}\left\{n_k(\tau)\right\}, && k = 1,2,\ldots,N \end{aligned}$$

The next theorem derives those statistics:

THEOREM 3. *Let μ_N be given as in Thm. 2, then:*

$$\begin{aligned} \nu_N(k) &= a(k)\frac{c_{N-k}}{c_N}, \\ \varsigma_N(k,l) &= a(k)a(l)\left(\frac{c_{N-k-l}}{c_N} - \frac{c_{N-k}c_{N-l}}{c_N^2}\right),\ k \neq l, \\ \sigma_N^2(k) &= a^2(k)\left(\frac{c_{N-2k}}{c_N} - \frac{c_{N-k}^2}{c_N^2}\right) + a(k)\frac{c_{N-k}}{c_N}, \end{aligned}$$

for $k,l = 1,\ldots,N$, and $c_{-m} = 0,\ m = 1,\ldots$.

Theorem 3 gives a characterization of the distribution of cluster sizes. The correlation between cluster sizes is resulting from the finite value of N and the constraint given in (1). In order to complete the characterization, we need to obtain the values $\{c_m\}$. These values can be derived using the series $S(x) = \sum_{i=1}^{\infty} a(i)x^i$ that is assumed to converge for $x \in D_S = \{x|\ |x| < R_S\}$.

THEOREM 4. *Under the assumption of convergence of the series $S(x)$ in D_S*

1. *The values $c_n,\ n = 1,2,\ldots$ are the coeffients of the Taylor expansion of the function $g(x) = e^{S(x)}$,* i.e.,

$$g(x) = e^{S(x)} = \sum_{n=0}^{\infty} c_n x^n$$

where $g(x)$ converges over $D_g = D_S$.

2. *The radius of convergence of the Taylor series of $g(x)$ and $S(x)$ are equal,* i.e.,

$$\lim_{n\to\infty} \frac{c_n}{c_{n+1}} = \lim_{n\to\infty} \frac{a_n}{a_{n+1}} = R.$$

3. *The values c_n can be derived by the recurrence relation: $c_0 = 1$, $c_1 = a(1)$,*

$$(n+1)c_{n+1} = \sum_{k=0}^{n} (k+1)a(k+1)c_{n-k},\ n = 1,2,\ldots$$

The above theorem gives a simple and efficient method to derive the value of c_N that are used to compute the distribution of cluster size based on theorem 3. The method can be summarized as:

1. Write the series $S(x)$ and obtain the function that it converges to.
2. By deriving the function $g(x) = e^{S(x)}$, find a recurrence equation relating the coefficient c_n.
3. Using the recurrence equation, the values c_n are obtained, yielding the distribution of cluster sizes.

We thus have a complete theoretical characterization of the distribution of cluster size for finite values of the number of nodes.

2.2 Mean field analysis

The above analysis enables us to derive an analytic description of merge–split systems with finite number of nodes. However, the procedure, even if straightforward, becomes imprecise, when the number of nodes N becomes large. This is because the values of c_n grows almost exponentially, and even for a 70 to 80 nodes, reach huge values close to 10^{71}. Doing addition as needed by the recurrence equation needed for the exact derivation with such large values leads to errors that propagates to all values. For this reason we need an approximation that is more amenable to calculation for systems with more than 70 nodes as is needed in practice. Moreover it does not give analytic insight into the large-scale behavior of systems based on merge–split processes. In order to deal with these two issues, we will present here an asymptotic analysis of merge–split processes, *i.e.*, the limit process when the number of nodes grows, $N \to \infty$. The asymptotic behavior of $\nu_N(k)$ is obtained through the next theorem:

THEOREM 5. *Suppose that*

$$R = \lim_{n\to\infty} \frac{c_n}{c_{n+1}},$$

then for fixed k, we have:

$$\begin{aligned} \lim_{N\to\infty} \nu_N(k) &= a(k)R^k, k = 1, 2, \dots \\ \lim_{N\to\infty} \varsigma_N(k,l) &= 0, k \neq l = 1, 2, \dots \\ \lim_{N\to\infty} \sigma_N^2(k,l) &= a(k)R^k, k = 1, 2, \dots \end{aligned}$$

This asymptotic behavior yields an interesting insight: it proves the existence of a limit when the number of nodes diverges, and it shows that the correlation between the number of clusters of different sizes vanishes with increasing number of nodes. This last property is called the propagation of chaos in the literature as it means that the correlation between states vanishes when the number of nodes diverges.

However, the above theorem is difficult to apply to systems with finite number of nodes because the physical conditions of the system change with the number of nodes. To keep node density constant as the number of nodes diverges, we proceed as follows. Assume that a finite system with n nodes is evolving in a unit volume. To maintain the same physical conditions, we let $V(n)$ grow along with the number of nodes n diverging, ensuring that the node density remains equal to N, *i.e.*, we are analyzing $\eta_N(k)$, the density of clusters of size k, when the node density is equal to N:

$$\lim_{n\to\infty, n=N.V(n)} \nu_n(k) = \eta_N(k),$$

where the constraint $n = N.V(n)$ results from the node density being N and subscript N indicates this. By extension $\eta(k,t)$ is defined as the density at time t of clusters of size k. We also define the merging rate per volume unit as

$$K_N(k,l) = \lim_{n\to\infty, n=N.V(n)} \frac{\psi_n(k,l)}{V(n)}.$$

Similarly we define the splitting rate per volume unit as

$$F_N(k|l) = \lim_{n\to\infty, n=N.V(n)} \frac{\phi_n(k|l)}{V(n)}.$$

Again, we will drop index N when it is obvious from the context. Recently, [3] proposed a methodical approach to derive the mean field of a special class of processes where the intensity of reactions in a system vanishes when the number of interacting nodes (N) increases. For such processes, the state occupation measure of the process converges to a mean-field limit that is given as the solution of the drift equation of the state occupation measure. Unfortunately, the intensity of the above merge–split process is not vanishing as we assume that asymptotically the density of merge and split reactions converge to $F(k|l) > 0$ and $K(k,l) > 0$. So the framework defined in [3] is not applicable here and we have to derive the mean-field directly. We will do this next.

Using the above notations and withThm. 5, which states that the correlation between the number of clusters $\varsigma_N(k,l) \to 0$ vanishes when $N \to \infty$, one can write the Kolmogorov forward equation of the Markov chain governing the merge–split process with infinite number of nodes as:

$$\begin{aligned} \frac{\partial \eta(m,t)}{\partial t} = \frac{1}{2}\sum_{l=1}^{m-1} K(l, m-l)\cdot\eta(l,t)\eta(m-l,t) \\ - \sum_{l=1}^{\infty} K(m,l)\cdot\eta(m,t)\eta(l,t) \\ + \sum_{k=m+1}^{\infty} F(k|m)\cdot\eta(k,t) \\ - \frac{1}{2}\sum_{l=1}^{m-1} F(m|l)\cdot\eta(m,t). \quad (5) \end{aligned}$$

We assume that the initial state $\eta(m,0),\ m = 1, \dots, \infty$, satisfies the node conservation condition, *i.e.*, $\sum_{k\geq 1} k\eta(k,0) = N$, the density of nodes being equal to N.

As in the discrete case, there are two special cases of this process. If one discards the split reaction, *i.e.*, $F(k|l) = 0$

for all $k < l$, one obtains a purely coalescent equation called Smoluchowski equation [23]. Discarding the merge reaction, *i.e.*, $K(k,l) = 0$ for all k, l, yields a purely branching process. The Smoluchowski equation has attracted an important and historical interest in the statistical physics community [9] because a considerable number of real world scenarios, *e.g.*, polymer synthesis in chemistry, aerosol formation in atmospheric studies, or phase separation in liquid mixtures, can be analyzed by solving it. More recently through the seminal survey by D. J. Aldous [1], the problem garnered increased interest in the mathematical community.

Now let us assume that cluster sizes are continuous, *i.e.*, $\upsilon(x,t)$ being the density of clusters of size x at time t, x being a continuous value. Then the Smoluchowski equation (Kolmogorov forward equation) in (5) governing the merge–split process turns into an integro-differential equation:

$$\begin{aligned}\frac{\partial \upsilon(x,t)}{\partial t} = \frac{1}{2}\int_0^x K(y,x-y)\upsilon(x-y,t)\upsilon(y,t)dy \\ - \int_0^\infty K(x,y)\upsilon(x,t)\upsilon(y,t)dy \\ + \int_0^\infty F(x+y|y)\cdot\upsilon(x+y,t)dy \\ - \frac{1}{2}\int_0^x F(x|y)\upsilon(y,t)dy. \quad (6)\end{aligned}$$

We are interested in deriving, when it exists, the asymptotic value $\upsilon(x) = \lim_{t\to\infty}\upsilon(x,t)$. When such an asymptotic value exists, it is the mean field approximation of the stationary distribution.

Fortunately, when the process is reversible, the stationary solution of the above integro-differential equation has a simple form that is given in the next theorem.

THEOREM 6. *The unique stationary solution $\upsilon(x) = \upsilon(x,\infty)$ of* (6) *for a reversible Markov chain satisfying the node conservation condition is:*

$$\upsilon(x) = a(x)e^{\lambda x}, \quad (7)$$

where λ is obtained subject to the node conservation condition (1)*,*

$$\sum_{k=1}^{\infty} k\upsilon(k) = N, \quad (8)$$

where N is the node density.

The next theorem shows the convergence of the system of N nodes to the mean field represented by the stationary solution given in (6) for a large class of merge–split processes.

THEOREM 7. *For all functions $a(x)$ satisfying $a(x) \sim x^\alpha e^{\gamma x}$ when $x \to \infty$, we have:*

$$\lim_{N\to\infty}\frac{\nu_N(k)}{\upsilon(k)} = 1.$$

This theorem yields a surprisingly simple large-scale behavior of the merge–split process that is called Mean Field Approximation (MFA). The MFA is of major interest as it provides a closed-form formula of the cluster size behavior relating $a(x)$, a microscopic parameter of the merge–split process, and through it the intensity ratio $q(x,y)$, to a macroscopic property of this process, the cluster size distribution $\upsilon(k)$. This closed-form function gives insight into the properties of the cluster size distribution that cannot be inferred easily by observing the exact distribution $\nu_N(k)$. In particular the MFA shows that the head of the distribution is controlled by $a(x)$, but the tail is determined by the exponents γ and λ, thus depending on the(finite) number of nodes.

Nonetheless, note that the convergence to the MFA is asymptotic. In particular, for large $k(N) < N$, the convergence of $\nu_N(k)$ to $a(k)R^k$ is known to be slow; *i.e.*, $\nu_N(k)$ and $\upsilon(k)$ might differ considerably for large $k(N) < N$.

2.2.1 Case study

To show the convergence to the mean field and the above described effects, we study two cases of interest: $a(i) = \beta$ and $a(i) = \frac{\beta}{i}$.

Case 1: $a(i) = \beta$**.** This is the case where the merge and split rates are constant and $q(i,j) = \frac{1}{\beta}$. The function $S(x)$ is derived as

$$S(x) = \sum_{i=1}^{\infty}\beta x^i = \frac{\beta x}{1-x},$$

with $D_S = (-1,1)$ and $g(x) = e^{\frac{\beta x}{1-x}}$. By deriving $g(x)$ we have $(1-x^2)g'(x) = \beta g(x)$, resulting in the below recurrence equations for $n = 1, 2, \ldots$:

$$c_0 = 1, c_1 = \beta, (n+1)c_{n+1} = (2n+\beta)c_n - (n-1)c_{n-1},$$

that leads to a monotonically increasing sequence $c_n,\ n > 0$. Therefore $\nu_N(k)$ is monodically decreasing with k, $1 \le k < N$ (it might increase for $k = N$).

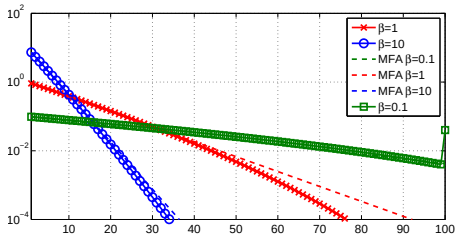
Applying the mean-field formula given in Theorem 6, with $a(x) = \beta$ we obtain $\lambda(N) = -\sqrt{\frac{\beta}{N}}$:

$$\upsilon(x) = \beta e^{-\sqrt{\frac{\beta}{N}}k}, \quad (9)$$

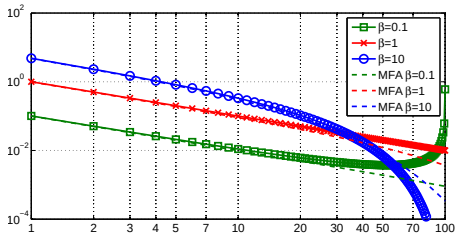
showing an exponential decrease of the number of clusters with the cluster size. The asymptotic distribution predicted by Thm. 5 is derived by noting that $R = \lim_{k\to\infty}\frac{a_k}{a_{k+1}} = 1$ and $\lim_{N\to\infty}\nu_N(k) = \beta$. Moreover $\lim_{N\to\infty}\lambda(N) = 0$, showing that $lim_{N\to\infty}\upsilon_N(k) = lim_{N\to\infty}\nu_N(k) = \beta$.

In Fig. 1a we show the distribution of cluster sizes obtained by the method described for finite number of nodes for a 100 nodes scenario, as well as the MFA given in 6. This demonstrates the remarkable quality of the mean field approximation, at least for small values of cluster sizes. Figure 1a also shows the loss of precision of the MFA for large k for finite value of N. These observations are in line with the theoretical analysis that predicted the MFA to be looser for large cluster sizes.

Case 2: $a(i) = \frac{\beta}{i}$**.** In this case $q(i,j) = \frac{ij}{\beta(i+j)}$. Such a function $a(i)$ can be used when clusters merge with a rate



(a) $a(i) = \beta$; note the logarithmic y axis.



(b) $a(i) = \frac{\beta}{i}$; note the logarithmic x and y axes.

Figure 1: Distribution of cluster sizes, $\nu_{100}(k)$, for a scenario of 100 nodes and several values of β.

proportional to the product of their size and split with a rate proportional to their size. With this assumption, the function $S(x)$ is derived as

$$S(x) = \beta \sum_{i=1}^{\infty} \frac{x^i}{i} = -\beta \log(1-x),\ D_S = (-1, 1).$$

Consequently $g(x) = \frac{1}{(1-x)^\beta}$. This results in

$$c_n = \frac{\beta(\beta+1)\dots(\beta+n-1)}{n!} = \frac{\Gamma(n+\beta)}{\Gamma(\beta)\Gamma(n+1)},\ n = 0, 1, \dots \tag{10}$$

Applying Thm. 3 generates the statistics of cluster sizes in a straightforward way. In particular for $\beta = 1$ we have $c_n = 1$ resulting in $\nu_N(k) = \frac{\beta}{k}$, which is independent of N.

The MFA for $a(x) = \frac{\beta}{x}$ is obtained as

$$\upsilon(x) = \frac{\beta}{x} e^{-\frac{\beta}{N}x} \tag{11}$$

and the asymptotic distribution predicted by Thm. 5 becomes $\lim_{N\to\infty} \nu_N(k) = \frac{\beta}{k}$.

We show in Fig. 1b the distribution of cluster sizes obtained for a 100 nodes scenario with $a(i) = \frac{\beta}{i}$ as well as the relevant MFA. Here also the MFA results in a remarkable approximation for small to moderate values of cluster sizes. However the approximation becomes looser for large cluster size because of the accumulation effect of finite N. By comparing the figures 1a and 1b, it can be seen that large size clusters are more frequent with $a(i) = \frac{\beta}{i}$ than when $a(i) = \beta$. In particular, for $\beta = 0.01$ the distribution shows on average 0.73 clusters with size 100 and on average 27% of the nodes are in clusters with other sizes, *i.e.*, the distribution is concentrated on a single cluster with 100 nodes.

Analysis of the correlation structure of the number of clusters gives interesting insights for this case. We show in Fig. 2, the correlation factor $\frac{\varsigma_N(k,l)}{\sigma_N(k)\sigma_N(l)}$ obtained through Thm. 3 for different values of β when $a(i) = \frac{\beta}{i}$. For $\beta \leq 0.1$, we observe a relatively strong correlation between values $\nu_N(k)$ and $\nu_N(N-k)$ (the values on the antidiagonal). Moreover, there is also a strong correlation between $\nu_N(N)$ and all other $\nu_N(k)$ (last row and column of the correlation factor matrix). This means that there are frequent direct transition from clusters of size $k < N$ to cluster of size N. The correlation on the antidiagonal can be interpreted as resulting from this last fact; most transitions are $C_N \to C_{N-k} + C_k$ and $C_{N-k} + C_k \to C_N$, *i.e.*, the number of clusters of size k and $N-k$ are expected to be almost equal and this confirmed by observing the curves in Fig. 1b that shows an almost symmetric curve of $\nu_N(k)$. When β becomes closer to 1, other transitions also appears. Nevertheless, when $\beta < 1$, these reactions occur essentially for large cluster sizes.

For $\beta > 1$, the correlation structure changes and becomes concentrated on the upper left triangle and for small cluster sizes, which can be interpreted by observing that now most transitions involve small clusters and rarely large ones.

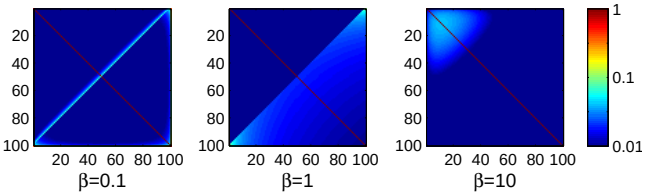


Figure 2: Correlation coefficient between number of clusters $\frac{\varsigma_{100}(k,l)}{\sigma_{100}(k)\sigma_{100}(l)}$ for 100 nodes obtained for $a(i) = \frac{\beta}{i}$ for different values of β, plotted with a logarithmic scale

2.3 Empirical fitting and parameters analysis

In practice we normally have access to microscopic information about the merge and split rate that results from the particular mobility pattern of a scenario. From these information one can estimate the intensity ratio $\hat{q}(i,j)$, that can be fitted to any functional form. However we saw previously that in order for the Markov process defined by the merge–split reactions to be reversible we should be able to find a function $a(i)$ such that $q(i,j) = \frac{a(i+j)}{a(i)a(j)}$. Moreover, Thm. 7, proving the convergence to the mean-field, suggests an asymptotic convergence to $a(i) = \beta\frac{e^{\gamma i}}{i^\alpha}$. Using such a functional form for $a(i)$ results in

$$q(i,j) = \frac{i^\alpha j^\alpha}{\beta(i+j)^\alpha},$$

showing that $q(i,j)$ does not depend on γ. Therefore α and β should be derived by fitting $q(i,j)$; γ can be estimated by applying the node conservation condition (1) on the exact cluster size distribution $\nu_N(k)$, *i.e.*, γ is chosen such that the resulting distribution $\nu_N(k)$ satisfies (1).

The parameters α and β are derived by fitting empirically derived values of intensity ratios to a function $\frac{i^\alpha j^\alpha}{\beta(i+j)^\alpha}$ by a

non-linear least-mean-square (LMS) technique. Frequently, the number of observed merge and split events becomes very small in particular for large cluster sizes, reducing their statistical value. A weighting equal to $\sqrt{m(i,j)s(i,j)}$ (where $m(i,j)$ is the number of merge events observed between clusters of size i and j and $s(i,j)$ is the number of split events of clusters of size $i+j$ to two clusters of sizes i and j, respectively) is applied to every measured intensity ratio. Moreover, for some cases the dynamical range of measured intensity ratio $q(i,j)$, is very large, *e.g.*, for small i and j, $q(i,j) \sim 0.001$ and for large i and j, $q(i,j) \sim 10$. In such cases we calibrate $\log q(i,j)$ to $\log \frac{i^\alpha j^\alpha}{\beta(i+j)^\alpha}$.

Knowing α, β and γ, the exponent λ to be used in the MFA can be obtained by solving the following equation:

$$\beta \sum_{k=1}^{N} \frac{e^{(\gamma+\lambda)k}}{k^{\alpha-1}} = N \tag{12}$$

The above fitting formulas give insight about the influence of the parameter values on the shape of the cluster size distribution. The curve of the cluster size distribution has two distinct parts: its head and its tail. The head of the distribution contains two essential pieces of information: the number of isolated nodes, *i.e.*, nodes that are not connected to any other nodes, and the slope of decrease of the distribution. Looking at the MFA, we can see that for small values of cluster sizes the distribution can be approximated as a polynomial with exponent $-\alpha$, and the number of isolated nodes is estimated as equal to $v(1) = \beta e^{\lambda}$, where λ depends on the number of nodes when α and β do not depend on it. The tail of the distribution is governed partly by the number of nodes that control directly γ and λ. Whenever $\lambda + \gamma$ becomes positive we can expect to see a bump on the tail of the distribution. This bump is the sign of emergence of a giant component (a well known phenomenon in the context of percolation theory [8]), as when N increases, the positiveness of the exponent $(\gamma + \lambda)$ will result in a greater number of large clusters.

The conditions under which such giant components emerge are determined by the value of $\lambda + \gamma$ that is controlled by (12). This equation states that if $\beta \sum_{k=1}^{N} \frac{1}{k^{\alpha-1}} < N$, $(\lambda + \gamma) > 0$ and one can expect to observe a bump in the tail of the distribution. For example when $\alpha = 1$, $\beta \sum_{k=1}^{N} \frac{1}{k^{\alpha-1}} = N$ and based on the value of β we have two qualitative behaviors: for $\beta < 1$, one observes a giant component, whereas for $\beta \geq 1$ no such component emerges. This is in line with the analysis given in Sec. 2.2.1, where $\beta = 1$ was found to be a boundary value for two types of behavior for the correlation structure of the finite system of nodes [1]. Indeed, the smaller the value of $\beta \sum_{k=1}^{N} \frac{1}{k^{\alpha-1}}$ (*i.e.*, the larger α and the smaller β), the larger the exponent $\lambda+\gamma$ will become and the stronger the tail bump will be and the larger the giant component. This last property helps understanding the meaning of the parameters and to interpret them in meaningful network properties: a large α and a small β (compared to N) means that the system of N nodes behaves in its stationary state as a small number (maybe even a single) giant component with single nodes splitting off or merging with it; a small α means that the network will remain an archipelago of disconnected clusters that merge and split.

[1] A similar analysis with stronger analytic basis can be done for the exponent γ alone (in place of $\lambda + \gamma$) and leads to mathematically stronger results but is omitted due to space restrictions

3. VALIDATION

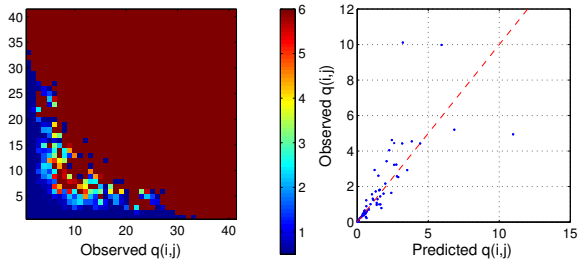
To this point, the analysis provided was strictly analytic. In this section we aim to validate that this mathematical analysis is of practical interest for predicting the behavior of realistic mobile networks. We will do this by analyzing a variety of scenarios: three real world scenarios as well as a synthetic random walk scenario. First, we will use the contact trace from Infocom 2005 as an example of a realistic mobile network and show that it can be described by a merge–split model. In the second part we study the random walk simulation, which serves to relate scenario parameters such as node density to the parameters of the merge-split process. Finally, we will analyze two large-scale traces based on GPS position traces from taxis in San Francisco and Shanghai to show the applicability of our model to real-world scenarios of hundreds and thousands of nodes.

3.1 Infocom 2005 contact data

In this subsection, we study the scenario described in [4]. In this experiment, 41 conference attendees of Infocom 2005 carried a small Bluetooth contact logger during the three days of the conference. Based on the Bluetooth contacts logged as tuples {device hardware address, contact start time, contact end time}, the connectivity graph has been reconstructed, allowing the merge and split rate function to be estimated empirically and their intensity ratios (defined in (2)) be derived. We plot the ratio $q(i,j)$ of those values in Fig. 3a: clearly, $q(i,j)$ increases with cluster sizes; nonetheless, a large part of the rate function remains undefined (shown with brown color relative to the NaN label in the figure) as no merge and split involving these values has been observed.

Applying the weighted least-mean-squares fitting described in Sec. 2.3 to the measured intensity ratio yields an estimation of $\hat{\alpha} = 3.71 \pm 0.1$, $\hat{\beta} = 16.73 \pm 0.95$ with a remarkable $R^2 = 0.998$ goodness of fit indicator. The value $\hat{\gamma} = 0.83$ is obtained by enforcing node conservation on the distribution $\nu_N(k)$. By enforcing node conservation on the MFA, one can derive $\lambda = -0.66$, resulting in $\lambda + \gamma = 0.17$ and therefore the emergence of a giant component. This can be verified by noting that $16.73 \sum_{k=1}^{41} \frac{1}{k^{3.71}} = 21.25 < 41$. In Fig. 3b, we plot the observed ratio $q(i,j)$ against the predicted ratio $\hat{q}(i,j) = \frac{a(i+j)}{a(i)a(j)}$, with $a(x) = \frac{16.73}{x^{3.71}e^{0.86x}}$.

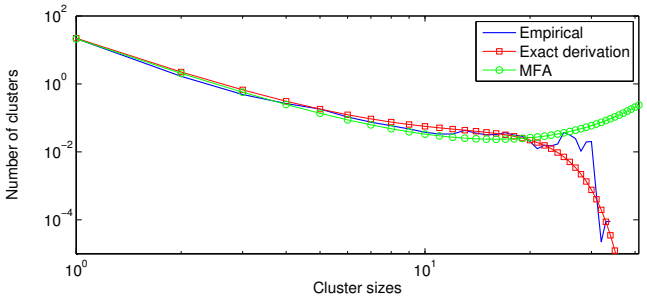
In Fig. 4a, we compare the cluster size vector observed over the entire trace with the distribution predictions introduced previously, *i.e.*, the exact derivation from Sec. 2.1, and the MFA from Sec. 2.2. The two distributions predict the empirical distribution with remarkable accuracy and the dif-



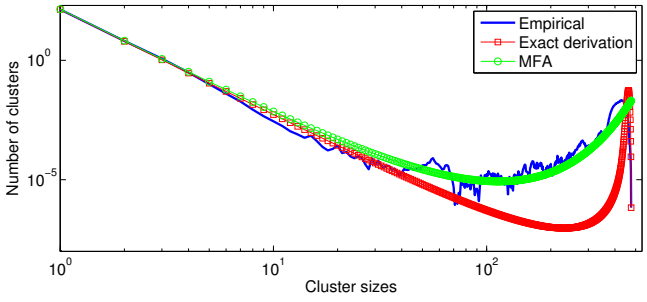
(a) Observed merge/split ratio $q(i,j)$

(b) Predicted $\hat{q}(i,j)$ *vs.* observed ratio $q(i,j)$

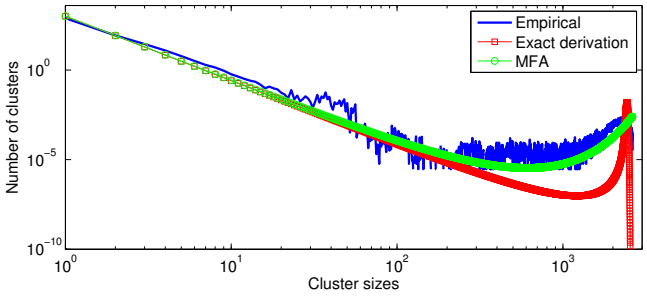
Figure 3: Merge/split ratio $q(i,j)$ and comparison with prediction in Infocom scenario



(a) Infocom



(b) San Francisco taxicabs



(c) Shanghai taxicabs

Figure 4: Empirical cluster size vector for real-world scenarios with exact derivation and MFA

ference between the MFA and the exact derivation are in line with the analysis provided previously. Note that even though the number of nodes is quite small, the exact derivation still yields a good prediction of the cluster size vector.

3.2 Synthetic random walk scenario

As a second scenario to validate our approach we used a synthetic random walk scenario. For this purpose, we run an extensive set of simulations with a simple home-grown mobility simulator that models mobile nodes moving according to the following random direction mobility model: at initial time $t = 0$, N nodes are placed uniformly at random in a square area. Then, each node is assigned a random direction in $[0, 2\pi)$. All nodes move in the assigned direction for l units, then they pick a new direction at random. If the trajectory of a node leads outside the simulation area it is reflected at the closest border. A link between two nodes is up if their Euclidean distance is less than the transmission range.

In Fig. 5, we plot the fitted values of α and β as a function of coverage (defined as the ratio between the area covered by the aggregated transmission range of all nodes and the simulation area); note that the coverage increases with the square of the transmission range. We observe that with

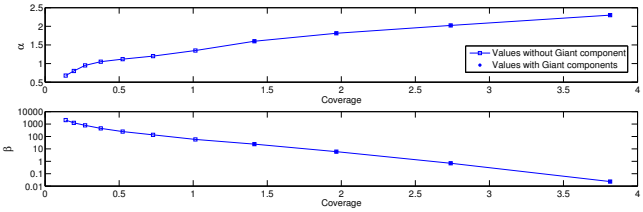


Figure 5: Estimated values of parameters α and β as a function of coverage for the random walk scenario

increasing coverage, α increases almost linearly and β decreases almost exponentially. Nevertheless, this holds only for coverage values above 1, where giant components may emerge. This observation is in line with what classical percolation theory predicts: giant components emerge only for high coverage. The figure also give us a benchmark about the value of α that one can expect to begin to observe large clusters emergence. The results of the cluster size vector fitting corresponding to the Infocom scenario is given in [2].

3.3 Taxicab Mobility Traces

We draw further statistics from two mobility traces based on GPS (Global Positioning System) position reports from taxicabs. Since our model is based on the adjacency matrix and yields the cluster size distribution, those positions reports cannot be used directly. Instead, in line with other recent publications [5], we assume that these taxies are equipped with some radio transmission technology (*e.g.*, IEEE 802.11) with a transmission range of 200 meters. This assumption affects the cluster size distribution from our model as well as the empirical one we compare it with in the same manner. In particular, a link between two nodes now means that those nodes are closer than 200m, yet this does not imply that those nodes would actually be able to communicate, as the range of wireless transmission depends on a great many number of factors, none of which are provided by the traces

we use. Further, since the GPS position reports contain a few dozen outliers, we use an MAD (Median of the Absolute Deviation) based filtering procedure [7] on the raw positions and remove these outliers. In order to increase the temporal resolution, we interpolate the position of cabs between reported positions but only if they are no further apart than a certain threshold in terms of distance and time. Therefore, the set of active nodes changes considerably over time. To reduce the effect of daily patterns, we limited the considered time range to 8AM until 12PM.

3.4 San Francisco Taxicab Mobility Trace

The traces from the San Francisco Cabspotting project have previously been studied in the context of DTN [20]; our trace contains 11.2 million GPS positions from 517 cabs.

We applied the model calibration over this data set and obtained estimates of $\hat{\alpha} = 4.437 \pm 0.004$ and $\hat{\beta} = 133.2 \pm 0.4$, $\hat{\gamma} = 0.3648$ with an $R^2 = 0.995$. The MFA was calibrated with a value $\hat{\lambda} = -0.3258$. The comparison of the empirical distribution of cluster sizes over the San Francisco taxis and the comparison with the exact derivation and the MFA are shown in Fig. 4b. This figure shows good agreement between the empirical distribution and the MFA. However, the quality of the prediction of the tail of the exact distribution degrades. This is to be expected as the number of nodes (being much larger than 70) yields c_n values beyond the limit of double precision floating point arithmetic. Indeed, for large scenarios the MFA can be the more suitable approximation as the figure shows. Of note, the San Francisco trace yields a larger value of α than the simulation scenario (Sec. 3.2), showing that in real scenarios nodes have a higher tendency to gather and build large clusters. Interestingly this tendency is even higher than for the Infocom scenario, where the exponent α is larger when the proportion of isolated nodes for the two scenarios are between 25% to 35%. This can be explained by the fact that taxis frequently gather at hot spots (train stations, restaurants, *etc.*), leading to a highly non-uniform distribution ([20] studies hot spots in this trace).

3.5 Shanghai Taxicab Mobility Trace

The Shanghai taxicab traces were collected by the Traffic Information Grid Team at Shanghai Jiaotong University [11]. The data consists of GPS position reports from 4063 taxies in Shanghai. The hot spots in this trace are studied in [14]. This trace contains even more nodes than the one from San Francisco as we monitored 3340 taxis. Here also the calibration of the ratio of intensity function is done and leads to $\hat{\alpha} = 3.602 \pm 0.1$ and $\hat{\beta} = 1007 \pm 3$ with an $R^2 = 0.9823$. The other parameters are also obtained as $\hat{\gamma} = 0.23$ and $\hat{\lambda} = -0.2242$. The comparison of the empirical distribution of cluster sizes over the Shanghai taxis trace and the comparison with the exact derivation and the MFA are shown in Fig. 4c. This figure shows very good agreement between the empirical distribution and the MFA. Here also as expected the exact derivation cannot give a good approximation as the number of nodes leads to computational artefacts. For this scenario, α is in the order of the Infocom scenario and smaller than the San Francisco scenario. The effect of the difference in α can be seen by observing that the bump in the tail of the San Francisco scenario is more pronounced than the one of the Shanghai scenario. The difference between the two taxicab scenarios might come from the differences in the gathering pattern of taxis and from the geographical and topographical difference between these two cities.

4. RELATED WORK

While routing in mobile ad hoc networks (MANETs) is based on the implicit assumption of connectivity, measurements from real mobile wireless networks (*e.g.*, [15, 12]) mooted this assumption to some extent. Due to the sparse nature of the scenarios for which measurement results could be obtained, those networks were found to be *disconnected* for the majority of time and sparked a new line of research, now with the opposite assumption of the network being disconnected. In those so-called delay-tolerant networks (DTN), single-hop communication opportunities called *contacts* are leveraged by forwarding algorithms to form over time so-called *space-time paths*. Studying in particular those space-time paths, [6] found a "small world" behavior in many mobility traces; [10] observes a "path explosion" phenomenon.

At a more abstract level, clustering has been found to be an important characteristic of mobile networks and algorithms building upon this property have been proposed, most of them complementing MANET routing with opportunistic forwarding between clusters (*e.g.*, [21, 18]). Furthermore, there are mobility models explicitly aiming to yield "realistic" clustering properties (*e.g.*, [18, 22]), motivated by the behavior of people.

Moreover, [20] studies mobility traces specifically in terms of clustering behavior with a focus on the relationship between cluster size and lifetime and introduces a heterogeneous random walk mobility model. This model is particularly interesting because it on purpose yields node behavior that makes them statistically indistinguishable and it also models clustering as a feature of the scenario, rather than as the result of some assumed social behavior of the nodes. Our work is similar in that we also characterize clustering without modeling any social behavior of nodes. Yet, instead of studying mobility we opt to derive the cluster size distribution as a *consequence* of mobility, with the benefit of obtaining simple, analytically tractable expressions.

In terms of methodology, [**?**] uses a similar mathematical approach for analyzing a network running a gossip protocol and they prove the emergence of a spatial mean field describing the age of the latest update received by mobile nodes .

Finally, the phenomenon of a phase transition for asymptotically large networks has been studied already in [19], and more recently been applied to asymptotically large mobile networks (see [8] and references therein).

5. CONCLUSION AND FUTURE WORK

Beginning with the observation that many real world mobile networks are partially connected, we develop a model for predicting the cluster size distribution in general systems of mobile nodes. The simple model we propose is of interest in particular because it yields a closed form result. The cluster size distribution is a metric that provides rich information as to the existence of partial paths, which enable faster forwarding schemes as opposed to purely contact-based algorithms. We prove that for large numbers of nodes, the proposed model converges to a mean field behavior, yielding a simple, closed form expression that translates the measurable merge and split behavior of clusters in a given scenario to the stationary cluster size distribution.

We validate the predictions from our model against a synthetic random walk mobility model and also with several real-world mobility traces ranging from tens to thousands of nodes in size. Motivated by the remarkable prediction quality for those traces, we believe that this model could be useful for studying several other questions.

- We analyze the transient behavior of the merge–split process, yielding further insight into the temporal characteristics of a scenario;
- we derive an upper bound on the fraction of nodes that can communicate in a disconnected network with a given delay bound;
- for an individual node, we derive the distribution of the size of its cluster after the subsequent merge event.

We hope that this paper serves as the basis of many more results furthering the understanding of the complex clustering phenomena of mobile networks.

6. REFERENCES

[1] D. J. Aldous. Deterministic and stochastic models for coalescence (aggregation, coagulation): a review of the mean-field theory for probabilists. *Bernoulli*, 5:3–48, 1997.

[2] Anonymous. Appendix to Globs in the Primordial Soup. Technical report. Download: `sites.google.com/site/edas1569285961/`.

[3] M. Benaim and J.-Y. Le Boudec. A class of mean field interaction models for computer and communication systems. *Performance Evaluation*, 65(11-12):823–838, 2008.

[4] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of Human Mobility on the Design of Opportunistic Forwarding Algorithms. In *INFOCOM '06*, Barcelona, Spain, April 2006.

[5] A. Chaintreau, J.-Y. Le Boudec, and N. Ristanovic. The age of gossip: spatial mean field regime. In *SIGMETRICS '09*, New York, NY, USA, 2009.

[6] A. Chaintreau, A. Mtibaa, L. Massoulie, and C. Diot. The diameter of opportunistic mobile networks. In *CoNEXT '07*, 2007.

[7] L. Davies and U. Gather. The identification of multiple outliers. *J. Amer. Stat. Assoc.*, 88(423):782–792, 1993.

[8] O. Dousse. *Asymptotic properties of wireless multi-hop networks*. EPFL Ph.D. Thesis no. 3310, 2005.

[9] R. L. Drake. *A General Mathematical Survey of the Coagulation Equation*, volume 2-3 of *International reviews in aerosol physics and chemistry*, pages 201–376. Pergamon Press, 1971.

[10] V. Erramilli, A. Chaintreau, M. Crovella, and C. Diot. Diversity of forwarding paths in pocket switched networks. In *IMC '07*, 2007.

[11] H.-Y. Huang, P.-E. Luo, M. Li, D. Li, X. Li, W. Shu, and M.-Y. Wu. Performance evaluation of SUVnet with real-time traffic data. *IEEE Transactions on Vehicular Technology*, 56(6):3381–3396, Nov. 2007.

[12] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and C. Diot. Pocket switched networks and human mobility in conference environments. In *WDTN '05*, Aug. 2005.

[13] F. P. Kelly. *Reversibility and Stochastic Networks*. Wiley, 1979.

[14] J. Lee, K. Lee, J. Jung, and S. Chong. Performance evaluation of a DTN as a city-wide infrastructure network. In *CFI '09*, 2009.

[15] V. Lenders, J. Wagner, and M. May. Analyzing the Impact of Mobility in Ad Hoc Networks. In *REALMAN 2006*, May 2006.

[16] A. A. Lushnikov. Coagulation in finite systems. *Journal of Colloid and Interface Science*, 65(2):276 – 285, 1978.

[17] A. H. Marcus. Stochastic coalescence. *Technometrics*, 10(1):133–143, Feb. 1968.

[18] M. Musolesi and C. Mascolo. Car: Context-aware adaptive routing for delay-tolerant mobile networks. *IEEE Transactions on Mobile Computing*, 8, 2008.

[19] T. Philips, S. Panwar, and A. Tantawi. Critical connectivity phenomena in multihop radio models. *IEEE Transactions on Information Theory*, 35(5):1044–1047, 1989.

[20] M. Piórkowski, N. Sarafijanovoc-Djukic, and M. Grossglauser. A Parsimonious Model of Mobile Partitioned Networks with Clustering. In *COMSNETS '09*, January 2009.

[21] N. Sarafijanovic-Djukic, M. Piórkowski, and M. Grossglauser. Island hopping: Efficient mobility-assisted forwarding in partitioned networks. In *IEEE SECON*, 2006.

[22] V. Srinivasan, M. Motani, and W. T. Ooi. Analysis and implications of student contact patterns derived from campus schedules. In *MobiCom '06*, 2006.

[23] M. von Smoluchowski. Drei Vorträge über Diffusion, Brownsche Molekularbewegung und Koagulation von Kolloidteilchen. *Phys. Z.*, 17:557–571 and 585–599, 1916.